INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Unit-Selection Attack Detection Based on Unfiltered Frequency-Domain Features

Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Christoph Busch

da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
{ulrich.scherhag, andreas.nautsch, christian.rathgeb, christoph.busch}@h-da.de

Abstract

Modern text-to-speech algorithms pose a vital threat to the security of speaker identification and verification (SIV) systems, in terms of subversive usage, i.e. generating presentation attacks. In order to distinguish between presentation attacks and bona fide authentication attempts, presentation attack detection (PAD) subsystems are of utmost importance. Until now, the vast majority of introduced spoofing countermeasures rely on speech production and perception based features. In this paper, we utilize the complete frequency band without further filter-bank processing in order to detect non-smooth transitions in the full and high frequency domain caused by unit-selection attacks. For the purpose of especially detecting unit selection attacks, the applicability of Fast Fourier Transformation (FFT) and Discrete Wavelet Transformation (DWT) is examined regarding non-smooth transitions in the full and high frequency domain, excluding filter-bank analyses. Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) classifiers are trained on the German Speech Data Corpus (GSDC) and validated on the standard ASVspoof 2015 corpus resulting in EERs of 7.1% and 11.7%, respectively. Despite language and data shifts, the proposed unit-selection PAD scheme achieves promising biometric performance and hence, introduces a new direction to voice PAD.

Index Terms: speaker recognition, presentation attack detection, spoofing, unit-selection, countermeasure

1. Introduction

Biometric systems can be seen as identity management systems. Subjects can be identified or verified by processing and comparing reference and probe voice samples. In security scenarios, the performance of a biometric system is examined regarding subversive usage with respect to different system levels [1]. Attacks at the sensor level are referred to as presentation attacks [2]. SIV systems are threatened in particular, due to the advanced development of speech synthesis techniques [3].

Voice presentation attacks are categorized in six attack types [4]: synthesis [5], voice conversion [4], mock-up [6], replay [7], unit-selection [3] and mimicry [8]. Figure 1 provides an overview on the different types of voice presentation attacks. In a speech synthesis attacks, attackers creates a synthetic voice of the targeted identity, in order to synthesize speech samples which are accepted by the SIV [5]. In a voice conversion attack, an existing speech sample of the impostor is altered, such that it becomes more similar to the voice signal of the target subject [4]. In a mock-up attack, the impostor generates a synthetic signal in order to circumvent SIV systems by causing high comparison scores without necessarily containing speech signals [6]. Replay attacks refer to the playback of a previous captured voice sample to the SIV system [7]. For unit-selection attacks, speech samples of the attacked subject are captured, segmented into parts, called units, and replayed in different sequence to the SIV system. Imitation or Mimicry is the attempt of an impostor to mimic an enrolled subject, in order to get access to the system via the foreign account [8].

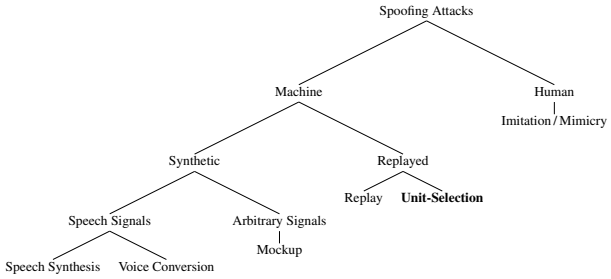


Figure 1: Structure of presentation attacks.

This paper is organized as follows: Section 2 provides a short overview over state-of-the-art PAD systems for unit-selection attacks. In section 3 unfiltered frequency-domain features are proposed and evaluated in section 4, followed by discussion and conclusion.

2. Voice PAD: ASVspoof 2015

In the ASVspoof 2015 spoofing challenge, the focus is placed on the detection of text-independent attacks, in particular synthesis, voice conversion, and unit-selection [3]. Five of ten attack algorithms are known during the PAD system development phase in order to investigate the detection robustness against unknown attacks. State-of-the-art voice PAD systems [9, 10, 11, 12] are capable of detecting 9 of 10 attack algorithms at EERs of about 0%. The countermeasures utilize phase-based features, which detect non-natural phase shifts in generated artefact samples, as during the synthesis process only amplitude information is concerned in the vocoding stage, making phase-based features convenient for detecting such artefacts. In contrast to synthetic speech signals, unit-selection creates artefacts by reusing previous recorded samples [13], thus the natural phase-shift of the sample is preserved and the applicability of countermeasures utilizing phase-based features is limited.

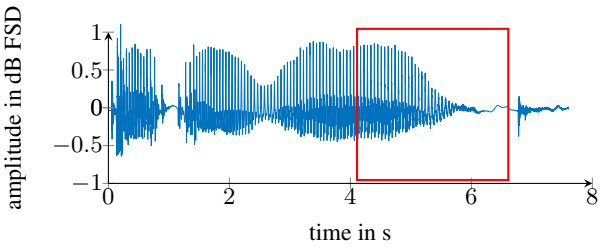
However, a successful countermeasure against unit-selection attacks, proposed in [14], employs a feature-combination of Cochlear Filter Cepstral Coefficients (CFCC), Instantaneous Frequency (IF), and Mel Frequency Cepstral Coefficient (MFCC). The CFCC, introduced in [15], is calculated by uti-

http://dx.doi.org/10.21437/Interspeech.2016-969

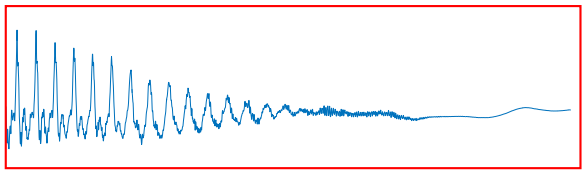
lizing an Audiotory Transform (AT), followed by a filter bank and Discrete Cosine Transform (DCT). The AT itself is a function emulating the filter function of the cochlear [16]. In order to take phase information into account, a CFCCIF is designed, combining CFCC with IF. Fused with MFCCs this approach yields an EER of 1.2% on the ASVspoof data, in particular an EER of 8.5% on unit-selection attacks, which is the best result achieved in the context of the ASVspoof 2015 [14]. In context of ICASSP 2016, the same authors proposed a unit-selection detection utilizing prosodic features, i.e. fundamental frequency (f_0) contour and strength of excitation (SoE), achieving an EER of 12.41% on the ASVspoof 2015 data [17].
Most common features that analyze frequencies, such as MFCCs of CFCCs, aim at emulating the perception of humans. However, the human hearing is rather specialized for speech recognition, thus state-of-the-art presentation attack countermeasures are capable of yielding significantly better PAD performances compared to human observers [18].

3. Frequency Analysis of Sound Unit Transitions

In our frequency-domain analysis of unit-selection attacks, speech is interpreted as a concatenation of phonemes or likewise sound units, where concatenation points are referred to as transitions. In bona fide (human) speech, the phonemes are smoothly transferred into each other. The continuous transition of a bona fide speech signal is depicted in figure 2.



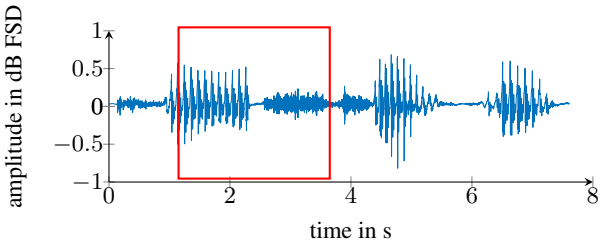
(a) Human speech signal.



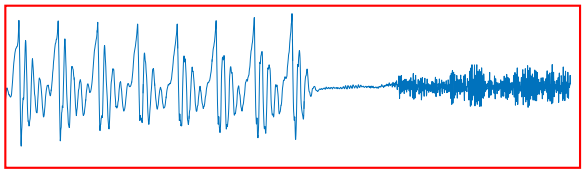
(b) Close-up of transition in human speech signal.

Figure 2: Example of a bona fide speech signal and transition.

Audio-signals which are compound of multiple voice fragments (phonemes, or other units) and not smoothed afterwards, show more abrupt changes of the frequency in the signal, as displayed in figure 3.
In bona fide speech, smooth transitions result in natural transitions in the frequency-domain as exemplary depicted in figure 4, whereas the transformation of the non-natural concatenated signal causes abrupt changes in the whole frequency band. Figure 5 illustrates unit-selection artefacts in the spectrum: in particular higher frequencies comprise abrupt changes in the magnitude, which compared to natural human speech comprise more density and occur more often. Motivated by this analysis, we propose Fourier- and wavelet-based features in order to distinguish between human speech and unit-selection attacks.

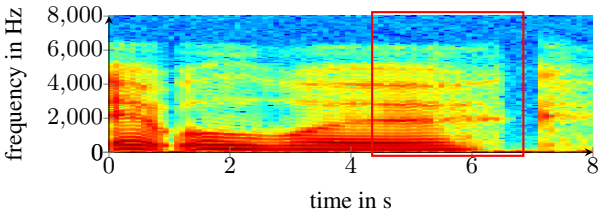


(a) Unit-selection speech signal.

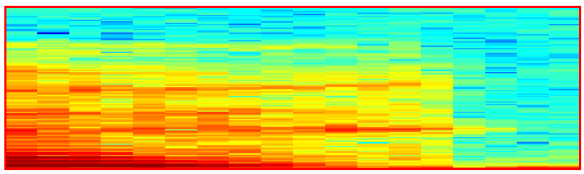


(b) Close-up of transition in unit-selection speech signal.

Figure 3: Example of a unit-selection speech signal and transition.



(a) Spectrogram of human speech signal.



(b) Close-up of spectrogram of transition in human speech signal.

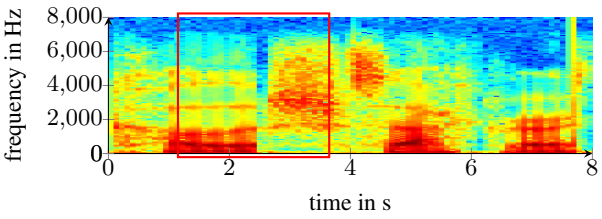
Figure 4: Spectrogram of a bona fide speech signal and transition.

3.1. Fourier-based Features

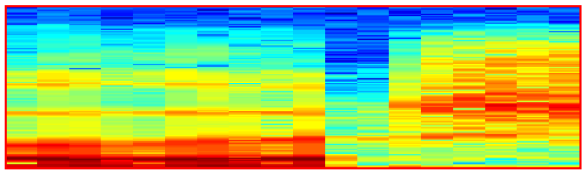
In contrast to the result of a Short Time Fourier Transform (STFT), as visualized in figure 5, the Fourier transform omits any time information. Thus, as the sudden changes in time domain, caused by non-natural transition, yield higher amplitudes for higher frequencies in frequency domain, a Fourier-based feature vector is motivated. The resulting vector of the Fourier transform represents the amplitude as natural value a and phase as imaginary values bi. For the purpose of compatibility with machine learning algorithms, the magnitude, $|a + bi|$ of the signal is calculated as: $|a + bi| = \sqrt{a^2 + b^2}$.

3.2. Wavelet-based Features

As visible in figure 5 and 4, higher frequencies are more significant for distinguishing bona fide from unit-selection samples than lower frequencies. A successive decomposition of a signal into bandpass-signals without losing information is possible, according to the Mallat theorem [19]. The Discrete Wavelet Transform (DWT) can be understood as bandpass filter decomposing the signal in iterative steps.



(a) Spectrogram of unit-selection speech signal.



(b) Close-up of spectrogram of transition in unit-selection speech signal.

Figure 5: Spectrogram of a unit-selection speech signal and transition.

Earlier iterations provide higher frequencies bands, later iterations lower. Assuming the discriminativity of higher frequencies, a feature vector extracting the fifth detail level is examined. This choice was elaborated based on experimental results employing 10 343 bona fide and 10 461 attack samples. In order to cover multiple frequency bands establishing more discriminative robustness, the proposed DTW feature comprises information fused from third to fifth iteration.
As the DWT represents a bandpass filter, the dimension of the result depends on the length of the analyzed signal. In order to obtain features with a fix dimension, a Fourier transformation is applied.

4. Experimental Results

The ASVspoof 2015 corpus [3] only contains unit-selection attacks in the evaluation set, thus we employ a contrastive database for development an optimization of countermeasures, in order to remain the ASVspoof data unseen, for the purpose of later comparison with countermeasures proposed at the spoofing challenge 2015. The unit-selection database is derived from the GSDC provided by the TU Darmstadt [20]. The unit-selection attacks are generated utilizing Mary-TTS [13]. The database protocol is depicted in table 1.

Table 1: Database partitioning on GSDC (self partitioned) and ASVspoof 2015 (S10 eval-set).

Subset	Bona Fide	Attack
Development Set	10 343	10 461
Calibration Set	3 745	4 484
Evaluation Set	400	100
ASVspoof S10	9 404	18 398

The proposed features are examined utilizing Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) as classifier, trained on the development set and optimized on the calibration set. The final performance of the classifiers is examined on the evaluation set. The performance of the countermeasures is assessed utilizing PAD subsystem metrics proposed by ISO/IEC CD2 30107-3 [2], in particular: Attack Presentation Classification Error Rate (APCER) and the Bona fide Classification Error Rate (BPCER). APCER is calculated as:

$$APCER = \frac{1}{N}\sum_{i=1}^{N}(1 - Res_i), \tag{1}$$

where N represents the number of attack presentations. Res_i takes value 1 if the i-th presentation is classified as an attack presentation, and value 0 if classified as a bona fide presentation. BPCER is calculated as:

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}}, \tag{2}$$

whereas N_{BF} represents the number of bona fide presentations. Res_i is defined as for the APCER. The EER of APCER and BPCER is estimated by the BOSARIS toolkit [21].

4.1. Model Training

The machine learning algorithms examined in this work are SVMs and GMMs. SVMs are chosen, as they represent a well-established machine learning algorithm which provides binary classification and are known for good pattern recognition performance [16]. Following the assumption, that Fourier based feature spaces comprise linear segregable populations, linear SVM kernels may yield adequate performance. As an alternative to the SVM approach, GMMs are trained. Log Likelihood Ratio (LLR) scores are conducted from 16-component GMMs representing bona fide and unit-selection speech, respectively. Assuming, that the proposed feature space results in different probabilistic clusters.

4.2. Results on Calibration Set

The EER of the machine learning algorithms strongly depends on the size of the analyzed feature vector. In an analysis of the frequency resolutions from 100 to 3000 in steps of 100, the most promising configurations are elaborated, depicted in table 2.

Table 2: Configuration for best EER.

Feature	Classifier	EER	Number of Frequencies
DWT-fusion+FFT	SVM	5.0%	600
	GMM	5.6%	200
FFT	SVM	6.1%	1000
	GMM	6.3%	1100
DWT-5+FFT	SVM	23.1%	100
	GMM	20.0%	1600

In general, a frequency resolution above 1100 bins leads to a rapid increasing EER. This effect is likely to be caused by the machine learning algorithms, as larger feature vectors require more training data in order to obtain satisfactory results.
On SVMs, the Fourier-based feature yields an EER of 6.1% on the calibration set with an FFT analyzing 1000 frequencies. The feature providing the fifth iteration of a wavelet fusion, referred to as DWT-5+FFT, yields an EER of 23.1%, a feature combining the third to fifth iteration, referred to as DWT-fusion+FFT, exceeds the basic Fourier approach by 1.1 percent points achieving 5.0%.

4.3. Results on Evaluation Set

The configurations scoring best on the calibration set are examined on the evaluation set. The observed EERs are depicted in table 3. In general, the EERs observed on the evaluation set are higher than on the calibration set. On this data set, the performance of the SVMs is less affected than those of the GMMs. The performance of the SVM analyzing the DWT-fusion+FFT-feature an EER of 7.1% is yielded, 2.1 percent points higher than for the calibration set.

Table 3: Best configurations evaluated with evaluation set and ASVspoof.

Feature	Comparator	EER Eval-set	EER ASVspoof
DWT-fusion+FFT	SVM	7.1%	11.7%
	GMM	15.0%	24.6%
FFT	SVM	8.5%	22.6%
	GMM	9.5%	27.7%
DWT-5+FFT	SVM	27.0%	11.7%
	GMM	40.1%	45.7%
CFCCIF [14]	GMM-UBM	–	8.5%

Figure 6 depicts the DET-plot for the examined algorithms. The DWT-5+FFT-feature is beyond the other features in all interesting operating points. Assessed with the SVM, the DWT-5+FFT-feature excels all other approaches for an APCER below 3%. The performance of FFT+SVM and DWT-fusion+FFT+SVM is approximately identical in most operation points, FFT+GMM is slightly inferior.

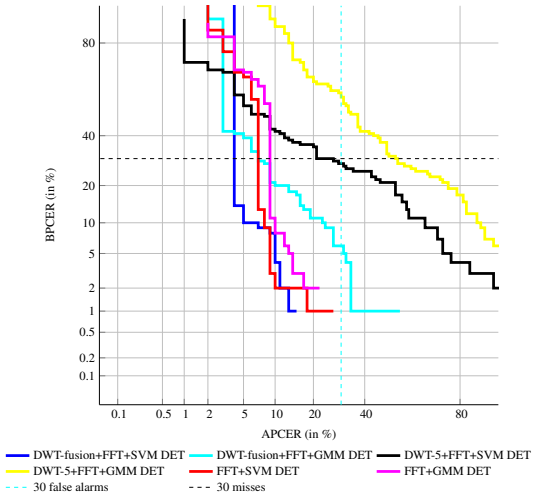


Figure 6: DET plots for configurations on evaluation set.

4.4. Results on ASVspoof data

Tested on the ASVspoof data, the performance of the proposed algorithms is slightly reduced, a comparison to the evaluation set is depicted in table 3. The EER of the DWT-fusion+FFT algorithm with SVM is raised by 4.6 percent points to 11.7%. Remarkable is the performance increase of the DWT-5+FFT feature analyzed, reducing the EER by 15.3 percent points to 11.7%. The DET of the DWT-5+FFT feature with SVM yields the best overall DET-plot on the ASVspoof data, as depicted in figure 7.

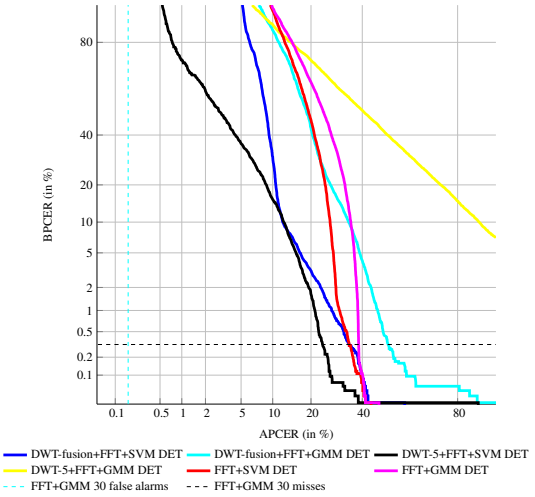


Figure 7: DET plots for configurations with best EER on ASVspoof S10 attacks.

5. Discussion

The proposed features are able to detect unit-selection attacks with an EER of 7.1% on the GSDC and 11.7% on the ASVspoof unit-selection attacks. Compared to the algorithms proposed at ASVspoof 2015, e.g. [14], the introduced features DWT-5+FFT (SVM) and DWT-fusion+FFT (SVM) yields competitive results with EERs of 11.7%, as depicted in table 3, operating on comparatively low computational costs, as a Fourier transformation (FFT) is utilized instead of the more expensive Spectrogram (STFT). Our proposed feature space and classifiers represent a contrastive PAD system, knowing the unit-selection attack scheme to face, which is unknown for the countermeasures described in section 2. However, our analysis comprise data shifts in terms of capture environments, the experimental setup, and the examined language. The field of PAD, especially unit-selection detection is very active. Current research proposes pitch analyses (fundamental frequency). A fusion of these features with low-level frequency analyses seems promising.

6. Conclusions

This paper shows that unfiltered frequency-domain features are feasible in PAD applications even when data and language shifts are persistent between development and evaluation data. Effective unit-selection voice PAD countermeasures are proposed by examining FFT and DWT properties of probe samples in a single-system fashion. Thereby, neither speech production nor speech perception theory is necessary. SVM- and GMM-classifiers are capable of distinguishing bona fide and unit-selection attacks samples. Further research on the frequency analysis of unfiltered speech signals promises improvement of the detection performance.

7. Acknowledgements

This work has been partially funded by the Center for Advanced Security Research Darmstadt (CASED) and the Hessen government (project no. 467/15-09, BioMobile)

8. References

[1] M. Faúndez-Zanuy, "On the vulnerability of biometric security systems," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 6, pp. 3–8, 2004.

[2] International Organization for Standardization, *Information Technology – Biometric presentation attack detection – Part 3: Testing and reporting*, JTC 1/SC 37 ISO/IEC 30 107-3:2016-CD2, 2016.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, C. Hanilc, A. Sizov, and U. Kingdom, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in *Proc. 16th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2015, pp. 2037–2041.

[4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, feb 2015.

[5] J. Villalba and E. Lleida, "Speaker Verification Performance Degradation against Spoofing and Tampering Attacks," in *Proc. Fala Workshop 2010*, 2010, pp. 131–134.

[6] C. Yang, G. Hammouri, and B. Sunar, "Voice Passwords Revisited," in *SECRYPT*, 2012, pp. 163–171.

[7] J. J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Proc. Carnahan Conference on Security Technology*. IEEE, oct 2011, pp. 1–8.

[8] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *Proc. 14th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2013, pp. 930–934.

[9] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: The NTU Approach for ASVspoof 2015 Challenge," in *Proc. 16th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2015, pp. 2052–2056.

[10] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative Phase Information for Detecting Human Speech and Spoofed Speech," in *Proc. 16th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2015, pp. 2092–2096.

[11] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, vol. 37, no. 2, 2016, pp. 1–5.

[12] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. Hansen, "Joint information from Nonlinear and linear features for spoofing detection: an i-vector/DNN based approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, vol. 53, no. 9, 2016, pp. 1689–1699.

[13] DFKI GmbH, "Unit selection voice creation and explanation on Individual Voice Import Components," 2014. [Online]. Available: https://github.com/marytts/marytts/wiki/UnitSelectionVoiceCreation

[14] T. B. Patel and H. A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech," in *Proc. 16th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2015, pp. 2062–2066.

[15] Q. Li and Y. Huang, "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification under Mismatched Conditions," *IEEE, Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, 2011.

[16] Q. Li, "An auditory-based transfrom for audio signal processing," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, no. 1. IEEE, oct 2009, pp. 181–184.

[17] T. B. Patel and H. A. Patil, "Effectiveness of Fundamental Frequency and Strength of Excitation for Spoofed Speech Detection," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, 2016, pp. 5105 – 5109.

[18] M. Wester, Z. Wu, and J. Yamagishi, "Human vs Machine Spoofing Detection on Wideband and Narrowband Data," in *Proc. 16th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2015, pp. 2047–2051.

[19] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, jul 1989.

[20] D. Schnelle-Walka, S. Radeck-arneth, C. Biemann, and S. Radomski, "An Open Source Corpus and Recording Software for Distant Speech Recogni- tion with the Microsoft Kinect," in *Proc. 11. ITG Fachtagung Sprachkommunikation*, 2014, pp. 1–4.

[21] N. Brümmer and E. D. Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing," Tech. Rep., 2011.